ĐƠN VỊ TỔ CHỨC

SỞ KHOA HỌC VÀ CÔNG NGHỆ THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐƠN VỊ PHỐI HỢP

SỞ GIÁO DỤC VÀ ĐÀO TẠO THÀNH PHỐ HỒ CHÍ MINH

THÀNH ĐOÀN THÀNH PHỐ HỒ CHÍ MINH

HỘI TIN HỌC THÀNH PHỐ HỒ CHÍ MINH

CÔNG TY TNHH MTV PHÁT TRIỂN CÔNG VIÊN PHẦN MỀM QUANG TRUNG

THƯỜNG TRỰC BTC

TRUNG TÂM PHÁT TRIỂN KHOA HỌC VÀ CÔNG NGHỆ TRẺ

TẬP HUẤN THÍ SINH DỰ THI

HỘI THI THỬ THÁCH TRÍ TUỆ NHÂN TẠO

THÀNH PHỐ HỒ CHÍ MINH NĂM 2026

AI CHALLENGE 2026

Kiến trúc Agentic AI và Ứng dụng mô hình ngôn ngữ lớn (LLM) trong hệ thống tìm kiếm

Hồ Lê Minh Quân
Trường Đại học Khoa học tự nhiên, ĐHQG-HCM

Table of contents

01 Các khái niệm cơ bản

02 Các nền tảng của AI Agent

03 Ứng dụng thực tế

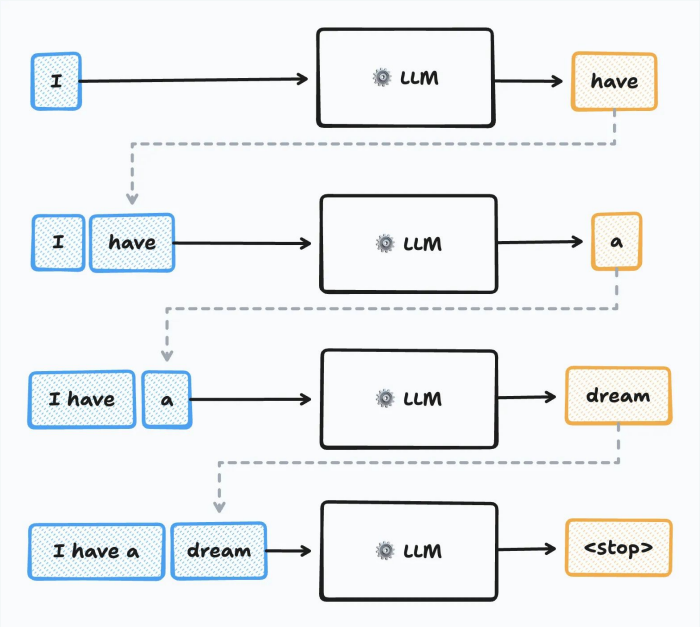
04 Kết luận

01

Các khái niệm cơ bản

Large Language Model (LLM)

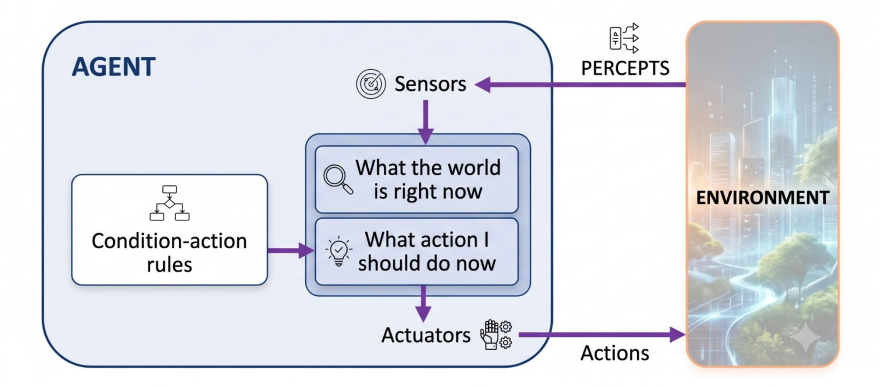
- Mô hình học sâu được huấn luyện trên lượng dữ liệu văn bản khổng lồ.
- Cơ chế: **Dự đoán token tiếp theo** dựa trên xác suất chuỗi trước đó.
- Khả năng: hiểu, xử lý ngữ nghĩa, tạo văn bản tự nhiên và thực hiện các tác vụ phức tạp (ví dụ: trả lời câu hỏi, hội thoại)
- Hạn chế: Phản xạ một chiều, phụ thuộc hoàn toàn vào luồng text đầu vào (prompt).



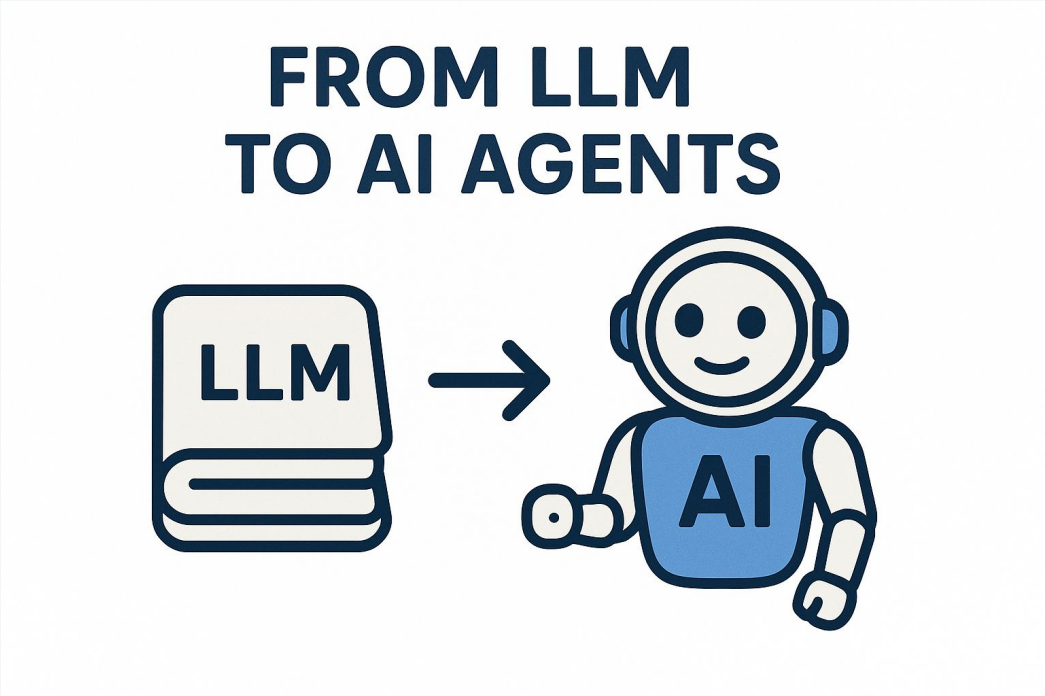
How LLMs Generate Text for the Rest of Us • pmdartus

AI Agent

- Tổng quát: Là hệ thống tự chủ có khả năng cảm nhận (**perceive**) thông qua các cảm biến và thực hiện hành động (**act**) lên môi trường để đạt mục tiêu.
- Thực tế, AI Agent đã tồn tại từ nhiều thập kỷ trước với các "bộ não" phi ngôn ngữ như:
  - Học tăng cường - VD: AlphaGo
  - Thuật toán Tìm kiếm - VD: Thuật toán A*)
  - Thị giác máy tính - VD: Xe tự lái.
- Hiện tại: Tích hợp LLM làm "bộ não" đang là **một hướng tiếp cận tốt nhất hiện nay** để xây dựng Agent (Language/LLM Agent).



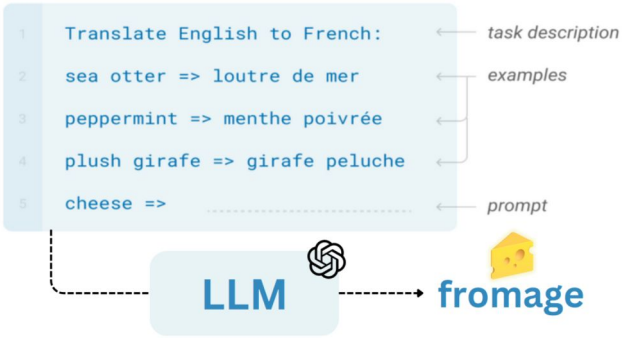
LLM vs AI Agent

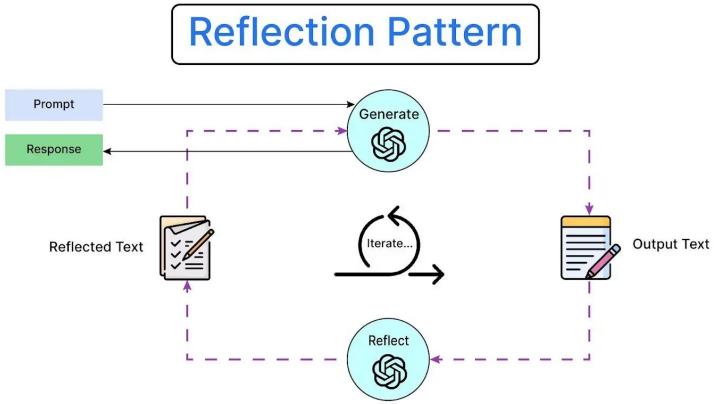


AI Agent + LLM

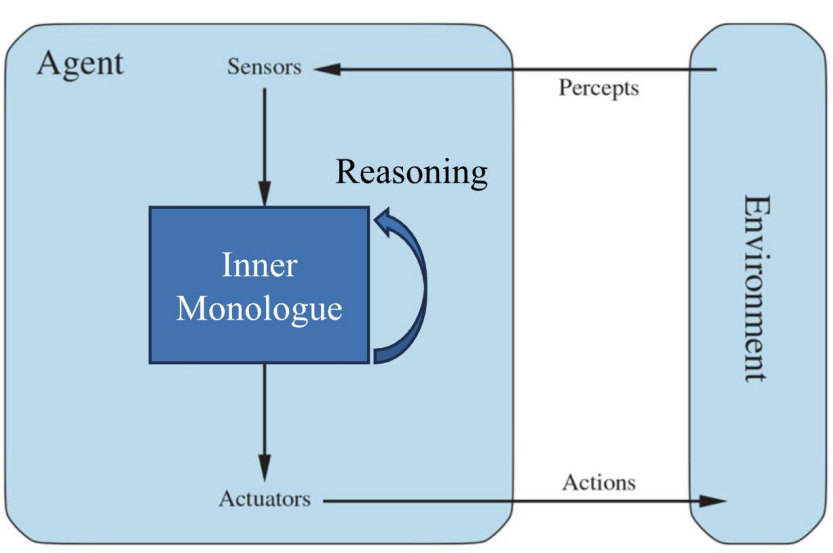
AI Agents khi được tích hợp LLMs ⟶ có thể ***sử dụng ngôn ngữ như một phương tiện để suy luận và giao tiếp***

- **Thực hiện chỉ dẫn tốt hơn:** Hiểu yêu cầu, học trong ngữ cảnh (in-context learning), tùy biến đầu ra theo nhu cầu
- **Suy luận để hành động thông minh hơn:** Diễn giải trạng thái, tự phản tư (self-reflect), điều chỉnh và tái lập kế hoạch





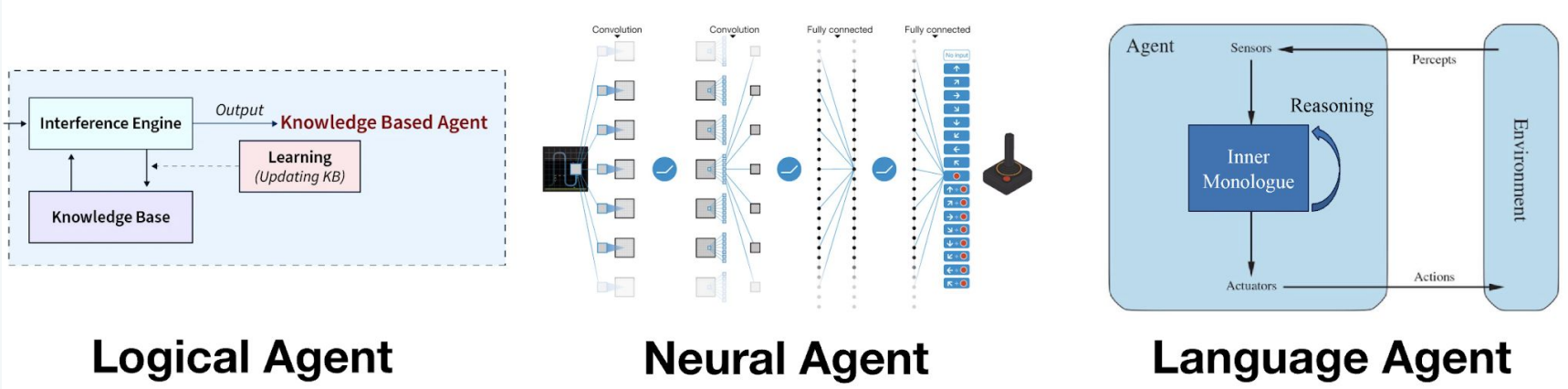
Action Space



Các khả năng của AI Agent khi tích hợp LLM:

- Internal Action
- Self-reflection
- Multimodal Perception

Sự phát triển của AI Agent



Knowledge-Based Agent in Artificial Intelligence - Scaler Topics

- ***Multimodal Agent***: agent đa phương thức (Thị giác, Âm thanh).
- ***LLM Agent:*** Lấy mô hình ngôn ngữ lớn làm lõi lập luận.

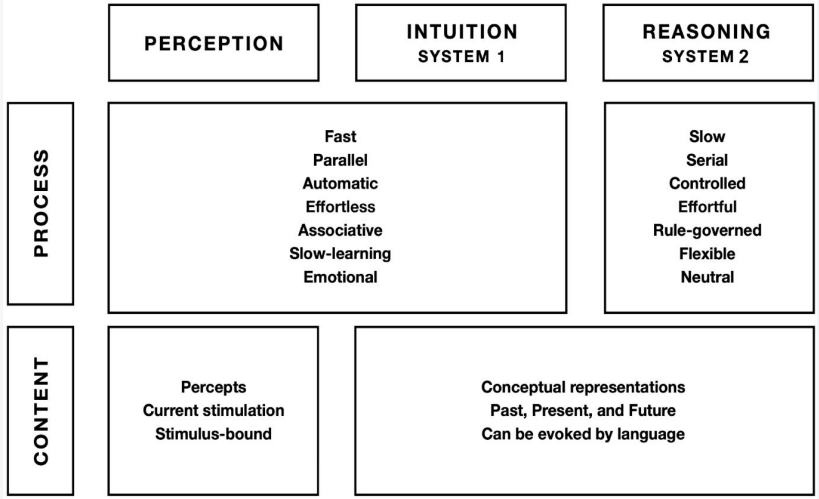
02

Các nền tảng của AI Agent

Reasoning

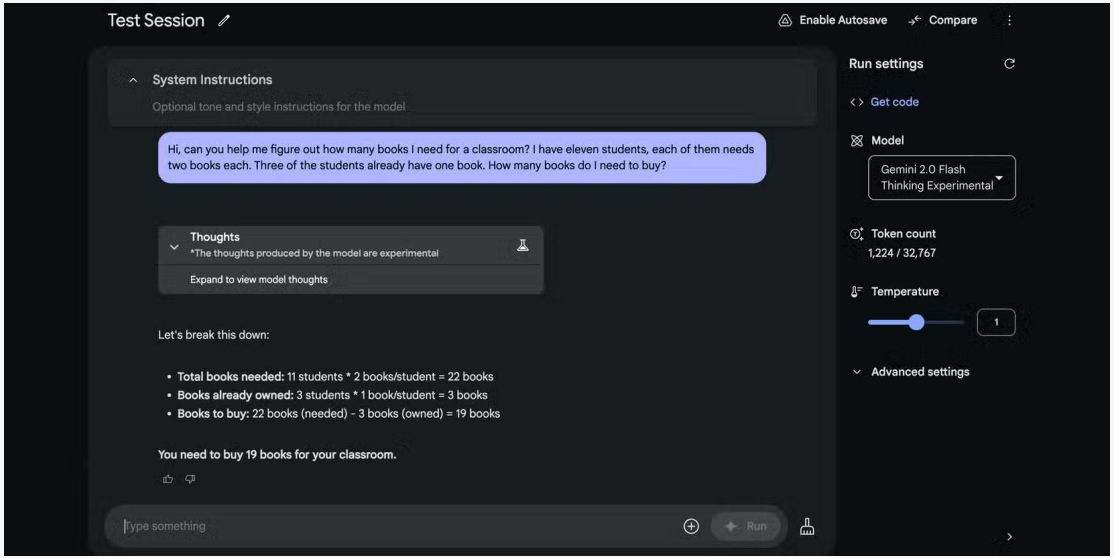
Cơ chế cốt lõi: Tạo sinh chuỗi token từng bước.

- **Hệ thống 1 (Trực giác - Intuitive):** Suy nghĩ nhanh, song song, phản xạ tự động. Dùng cho các tình huống ra quyết định nhanh.
- **Hệ thống 2 (Lý trí - Reasoning):** Chậm, tuần tự, kiểm soát. Cần nỗ lực để phân tích logic các bài toán phức tạp.



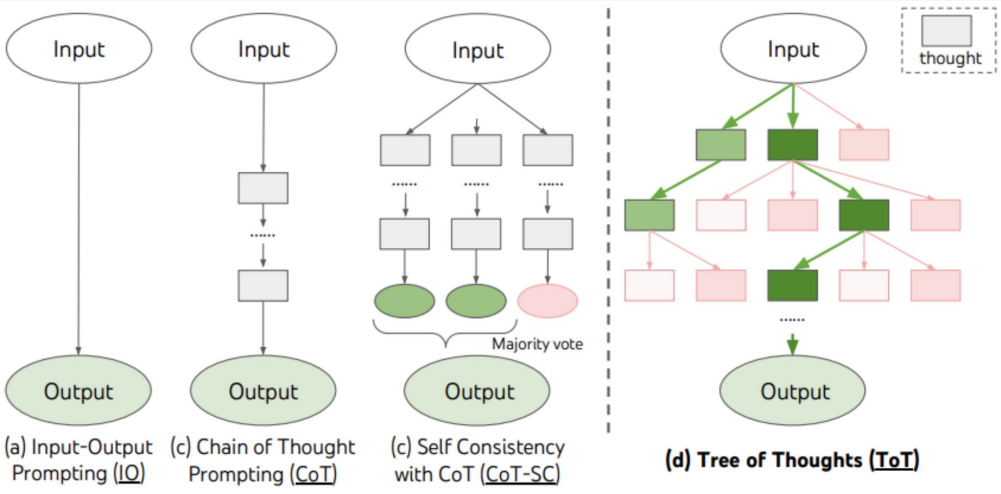
Reasoning

- *Trước đây:* LLM trả lời ngay lập tức (System 1), dễ mắc lỗi logic.
- *Hiện tại:* Tích hợp cơ chế **tư duy ngầm (Thinking)** trước khi xuất token cuối cùng.
- Giúp Agent lập luận vững chắc hơn trong toán học và truy xuất sự kiện phức tạp.
- Đại diện tiêu biểu: Dòng OpenAI o1, o3, Gemini 2.0 Flash Thinking.


Test Session
Enable Autosave
Compare
System Instructions
Optional tone and style instructions for the model
Hi, can you help me figure out how many books I need for a classroom? I have eleven students, each of them needs two books each. Three of the students already have one book. How many books do I need to buy?
Thoughts
*The thoughts produced by the model are experimental
Expand to view model thoughts
Let's break this down:
Total books needed: 11 students * 2 books/student = 22 books
Books already owned: 3 students * 1 book/student = 3 books
Books to buy: 22 books (needed) - 3 books (owned) = 19 books
You need to buy 19 books for your classroom.
Type something
Run
Run settings
Get code
Model
Gemini 2.0 Flash Thinking Experimental
Token count
1,224 / 32,767
Temperature
1
Advanced settings

Reasoning

- **Input-Output (IO):** Hỏi & Trả lời trực tiếp. Không có quá trình suy luận phức tạp, thiếu khả năng phân tích sâu.
- **Chain of Thought (CoT):** Suy luận theo một chuỗi đơn lẻ, giúp Agent suy nghĩ có hệ thống, nhưng vẫn có rủi ro nếu đi sai hướng.
- **Tree of Thoughts (ToT):** Tìm kiếm theo cây.
  - Tạo nhiều nhánh suy nghĩ (branches).
  - Tự đánh giá từng nhánh & Quay lui (backtracking) để chọn kết quả tối ưu nhất.



Memory - Phân loại

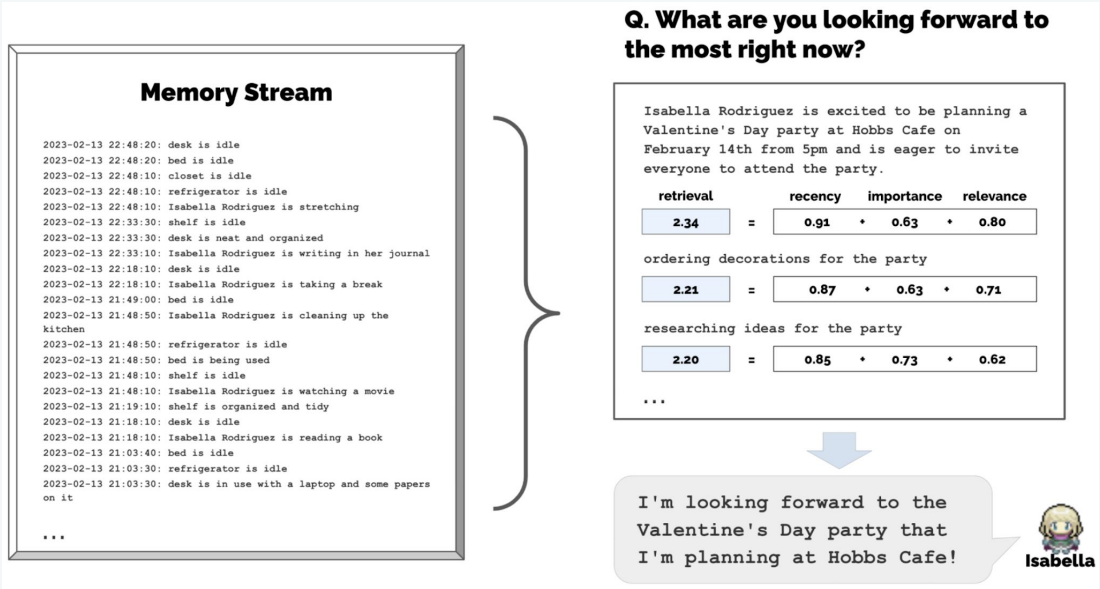
Loại theo nội dung	Định nghĩa	Ví dụ
Bộ nhớ theo tập (Episodic)	Lưu trữ trải nghiệm	Generative agents [Park et al., 2023]
Bộ nhớ ngữ nghĩa (Semantic)	Lưu trữ kiến thức	
Bộ nhớ thủ tục (Procedural)	Lưu trữ kỹ năng	Voyager [Wang et al., 2023]

Ghi chú: Ở đây phân loại dựa trên **nội dung bộ nhớ**, được:

- Lấy cảm hứng từ **hệ thống bộ nhớ dài hạn của con người**
- Độc lập với cách triển khai cụ thể (orthogonal to implementations)

Memory - Bộ nhớ hồi tưởng (episodic memory)

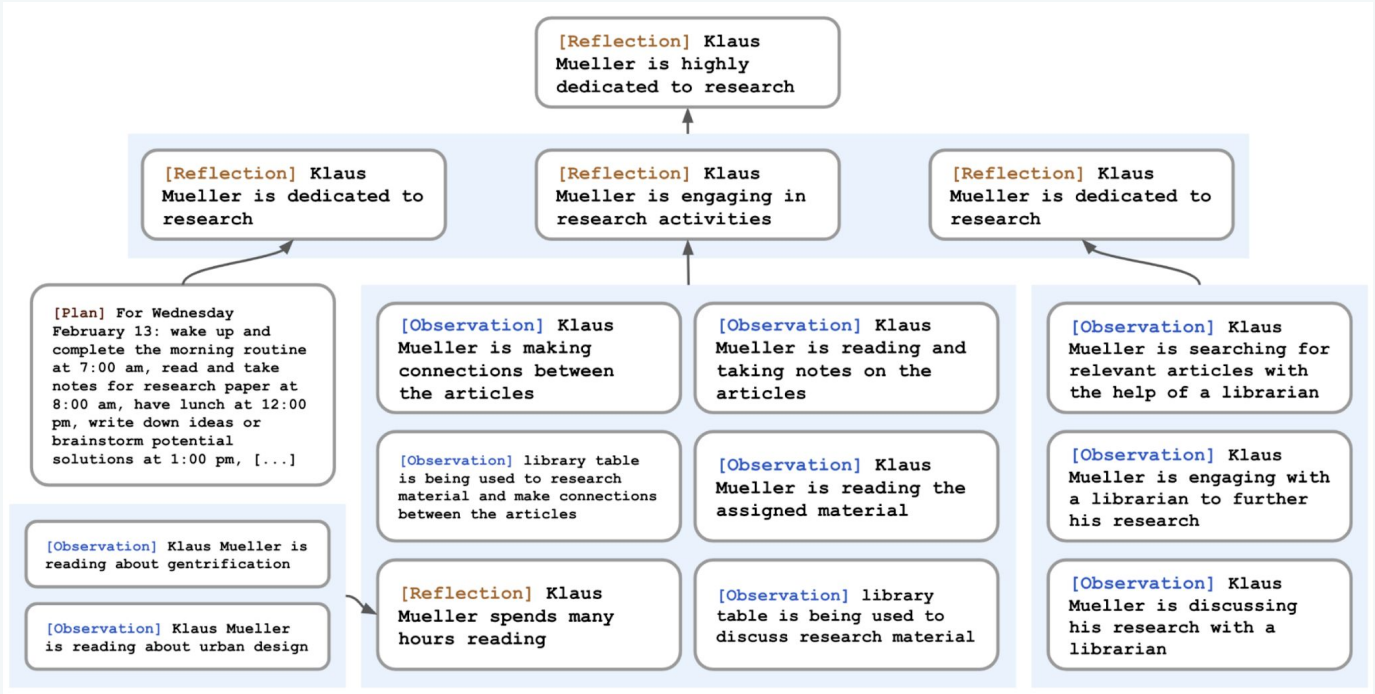
- Ghi nhớ (Write): chuỗi sự kiện ***chỉ được ghi nối tiếp*** (append-only event streams)
- Đọc lại (Read): truy xuất ***dựa trên điểm số heuristic*** (điểm đánh giá theo kinh nghiệm)


Memory Stream
2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10: desk is idle
2023-02-13 22:18:10: Isabella Rodriguez is taking a break
2023-02-13 21:49:00: bed is idle
2023-02-13 21:48:50: Isabella Rodriguez is cleaning up the kitchen
2023-02-13 21:48:50: refrigerator is idle
2023-02-13 21:48:50: bed is being used
2023-02-13 21:48:10: shelf is idle
2023-02-13 21:48:10: Isabella Rodriguez is watching a movie
2023-02-13 21:19:10: shelf is organized and tidy
2023-02-13 21:18:10: desk is idle
2023-02-13 21:18:10: Isabella Rodriguez is reading a book
2023-02-13 21:03:40: bed is idle
2023-02-13 21:03:30: refrigerator is idle
2023-02-13 21:03:30: desk is in use with a laptop and some papers on it
...
Q. What are you looking forward to the most right now?
Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.
retrieval
recency
importance
relevance
2.34 = 0.91 • 0.63 • 0.80
ordering decorations for the party
2.21 = 0.87 • 0.63 • 0.71
researching ideas for the party
2.20 = 0.85 • 0.73 • 0.62
...
I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!
Isabella

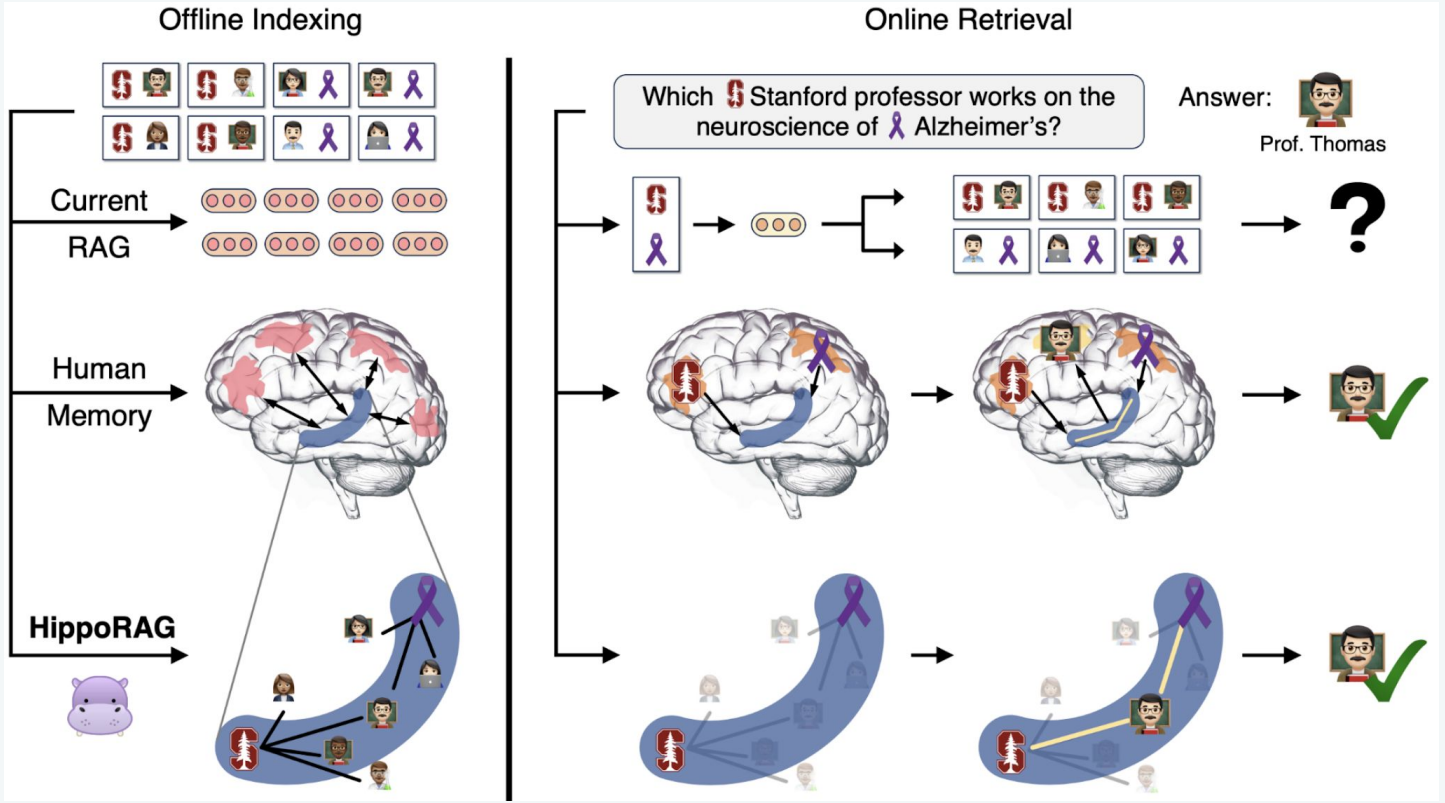
Tham khảo từ *EMNLP 2024 Tutorial on Language Agents*

Memory - Bộ nhớ ngữ nghĩa (semantic memory)

- Ghi nhớ (Write): LLM suy luận dựa trên các sự kiện
- Đọc lại (Read): truy xuất thông tin


[Reflection] Klaus Mueller is highly dedicated to research
[Reflection] Klaus Mueller is dedicated to research
[Reflection] Klaus Mueller is engaging in research activities
[Reflection] Klaus Mueller is dedicated to research
[Plan] For Wednesday February 13: wake up and complete the morning routine at 7:00 am, read and take notes for research paper at 8:00 am, have lunch at 12:00 pm, write down ideas or brainstorm potential solutions at 1:00 pm, [...]
[Observation] Klaus Mueller is making connections between the articles
[Observation] Klaus Mueller is reading and taking notes on the articles
[Observation] Klaus Mueller is searching for relevant articles with the help of a librarian
[Observation] library table is being used to research material and make connections between the articles
[Observation] Klaus Mueller is reading the assigned material
[Observation] Klaus Mueller is engaging with a librarian to further his research
[Observation] Klaus Mueller is reading about gentrification
[Observation] Klaus Mueller is reading about urban design
[Reflection] Klaus Mueller spends many hours reading
[Observation] library table is being used to discuss research material
[Observation] Klaus Mueller is discussing his research with a librarian

Memory - HippoRAG



Gutiérrez et al. *"HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models."* NeurIPS (2024)

Planning

Cho một mục tiêu G, xác định một chuỗi hành động $(a_0, a_1, \dots, a_n)$sao cho dẫn đến một trạng thái vượt qua bài kiểm tra mục tiêu $g(\cdot)$

- Xu hướng chung:
  - Tăng khả năng biểu đạt trong việc ***xác định mục tiêu***, VD: dùng ngôn ngữ tự nhiên thay vì ngôn ngữ hình thức
  - Không gian hành động (action space) được mở rộng hoặc ***không có giới hạn***
  - Tăng độ khó trong việc ***kiểm tra mục tiêu*** một cách tự động

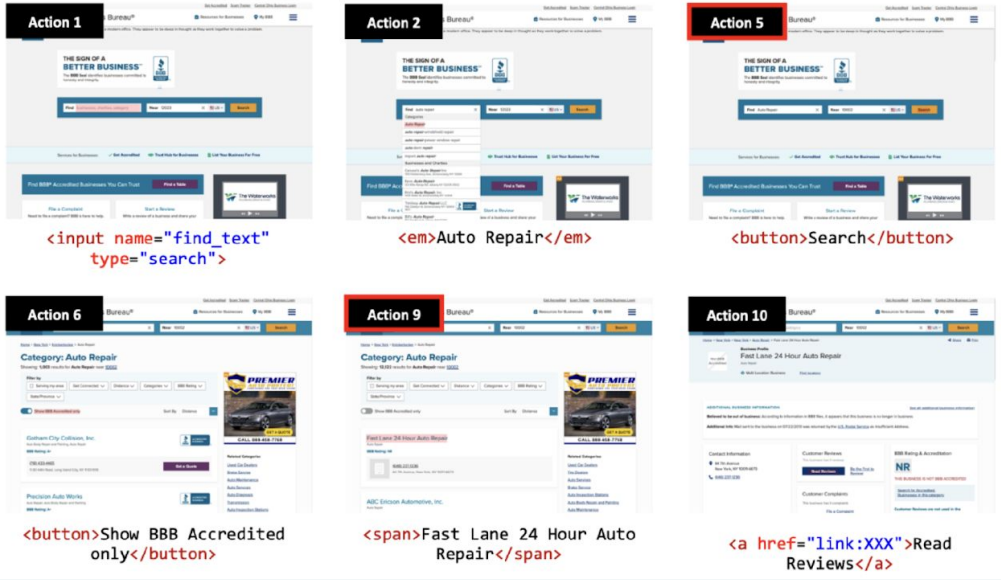
Planning

Task Description:
Show me the reviews for the auto repair business closest to 10002.

Action Sequence:

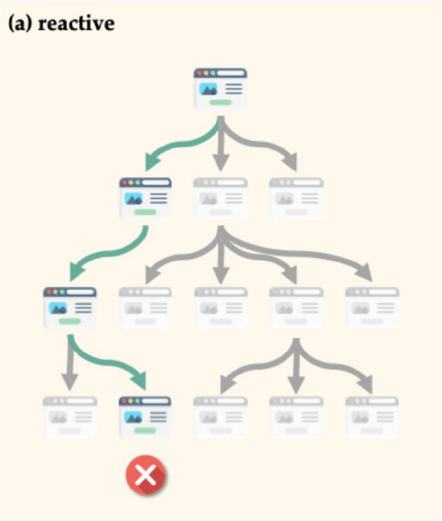
	Target Element	Operation
1.	[searchbox] *Find*	**TYPE: *auto repair***
2.	[button] *Auto Repair*	**CLICK**
3.	[textbox] *Near*	**TYPE: *10002***
4.	[button] *10002*	**CLICK**
5.	**[button] *Search***	**CLICK**
6.	[switch] *Show BBB Accredited only*	**CLICK**
7.	[svg]	**CLICK**
8.	[button] *Sort By*	**CLICK**
9.	**[link] *Fast Lane 24 Hour Auto Repair***	**CLICK**
10.	[link] *Read Reviews*	**CLICK**

Webpage Snapshots:

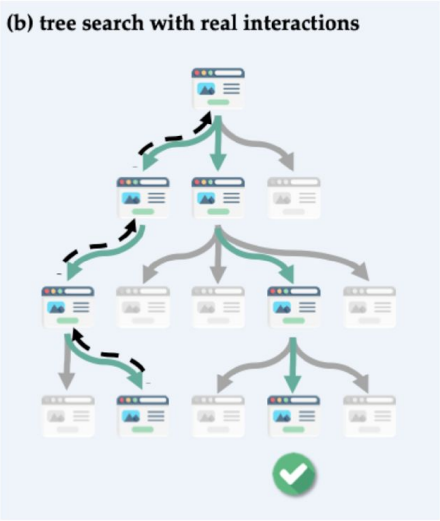


Deng et al., "Mind2Web: Towards a Generalist Agent for the Web." NeurIPS (2024)

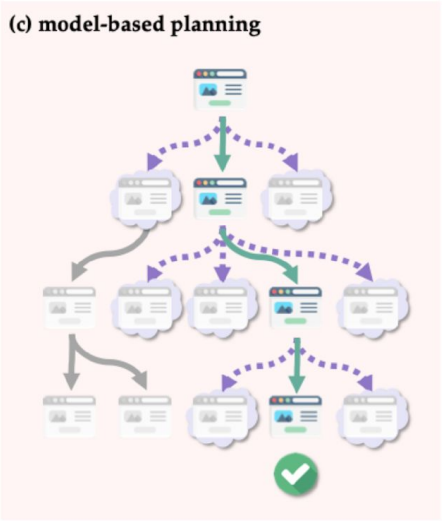
Planning



✅ Nhanh, dễ triển khai
⚠️ Tham lam, thiển cận



✅ Khám phá có hệ thống
⚠️ Hành động không thể đảo ngược, không an toàn, chậm



✅ Nhanh hơn, an toàn hơn, khám phá có hệ thống đảo ngược
⚠️ Làm thế nào để có được world model?

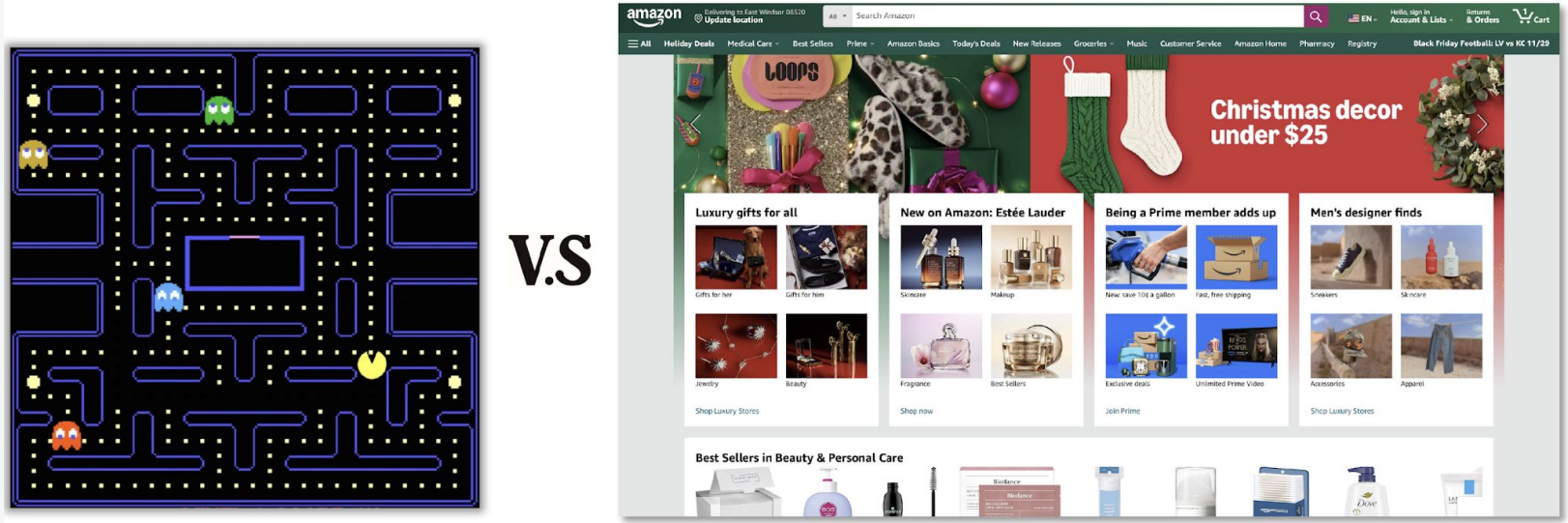
Planning - World Model

- Một mô hình tính toán của động lực chuyển trạng thái của môi trường
  - Ký hiệu toán học:

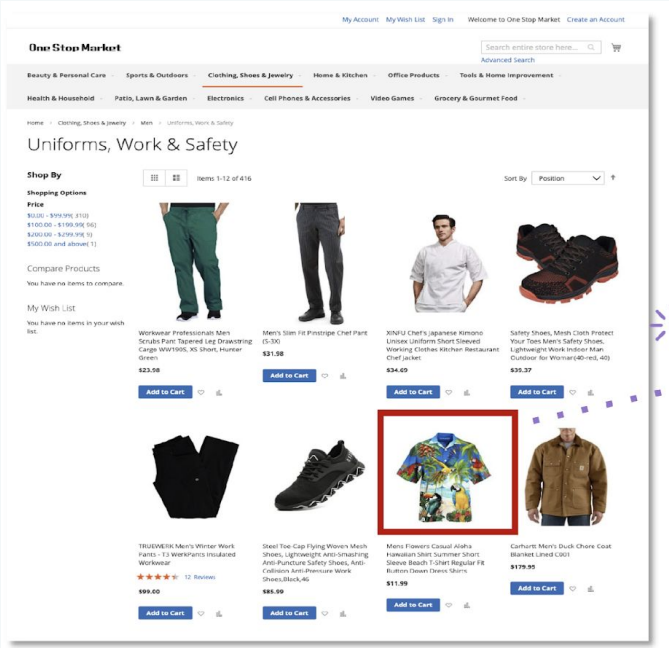
$$\hat{\mathrm{T}}: S \times A \rightarrow S$$

  - Mô hình nhận vào trạng thái (s_t) và hành động (a_t), cho ra trạng thái kế tiếp (s_{t+1})?

Planning - World Model



Planning - World Model



The page will navigate to a detailed product page for the "Mens Flowers Casual Aloha Hawaiian Shirt Summer Short Sleeve Beach T-Shirt Regular Fit Button Down Dress Shirts." This new page will likely contain additional information about the product including more detailed specifications, customer reviews, larger images, sizing options, and possibly a larger "Add to Cart" button. Other elements from the current category view like the grid of products will be replaced with the detailed view of this specific product.

Planning - World Model

Cook the potato and put it into the recycle bin.

LLM generates the high-level plan

Create a high-level plan for completing a household task using the allowed actions and visible objects.

Allowed actions: OpenObject, CloseObject, PickupObject, PutObject, ToggleObjectOn, ToggleObjectOff, SliceObject, Navigation

<In-context Examples>

Task description: Cook the potato and put it into the recycle bin.
Completed plans:
Visible objects are microwave, fridge, garbagecan, chair
Next Plans:

State

Action

Simulator

GPT3

High-level Plans

Plan: Navigation potato, PickupObject potato, ...

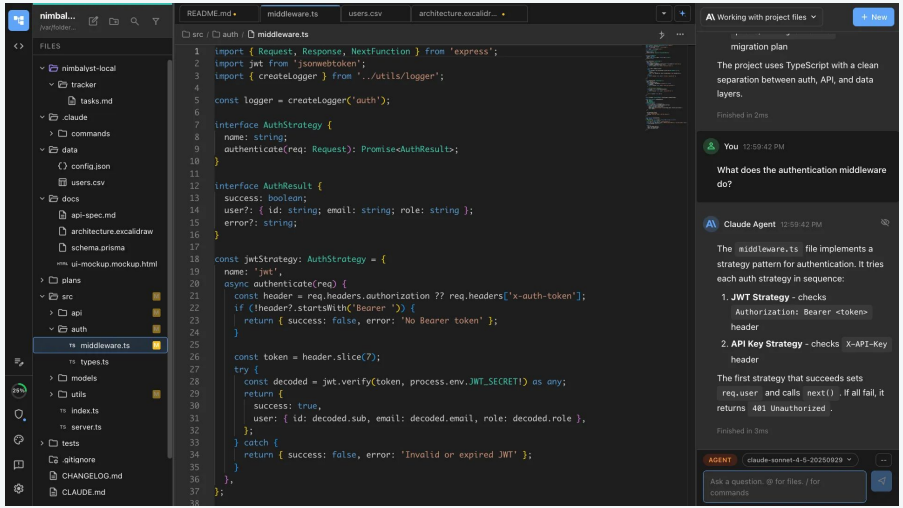
03

Ứng dụng thực tế

Ứng dụng AI Agent - Coding Agent

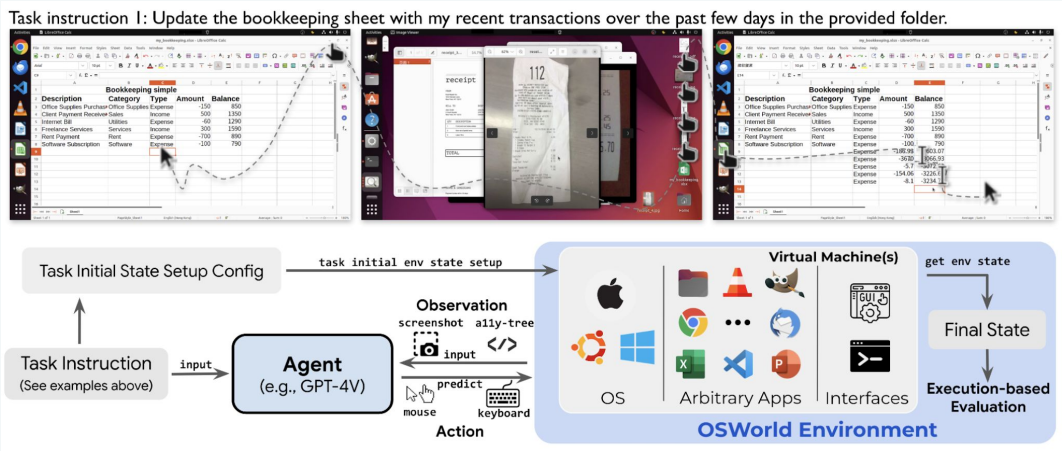
Nhiệm vụ lập trình cấp độ dự án

- **Agent:** Lõi suy luận lập trình tiên tiến (VD: Claude 3.5/3.7 hoạt động tự chủ qua các extension trên IDE).
- **Environment**: Code repositories, file systems, IDEs
- **Observation Space**: Source code, kết quả thực thi, tài liệu, thông báo lỗi, lịch sử commit,...
- **Action Space**: Chỉnh sửa mã, tìm/duyệt tệp, cập nhật kiểm thử (unit tests),...


README.md
middleware.ts
users.csv
architecture.excalidr...
src / auth / middleware.ts
import { Request, Response, NextFunction } from 'express';
import jwt from 'jsonwebtoken';
import { createLogger } from '../utils/logger';
const logger = createLogger('auth');
interface AuthStrategy {
name: string;
authenticate(req: Request): Promise<AuthResult>;
}
interface AuthResult {
success: boolean;
user?: { id: string; email: string; role: string };
error?: string;
}
const jwtStrategy: AuthStrategy = {
name: 'jwt',
async authenticate(req) {
const header = req.headers.authorization ?? req.headers['x-auth-token'];
if (!header?.startsWith('Bearer ')) {
return { success: false, error: 'No Bearer token' };
}
const token = header.slice(7);
try {
const decoded = jwt.verify(token, process.env.JWT_SECRET!) as any;
return {
success: true,
user: { id: decoded.sub, email: decoded.email, role: decoded.role },
};
} catch {
return { success: false, error: 'Invalid or expired JWT' };
}
},
};
Working with project files
New
migration plan
The project uses TypeScript with a clean separation between auth, API, and data layers.
Finished in 2ms
You 12:59:42 PM
What does the authentication middleware do?
Claude Agent 12:59:42 PM
The middleware.ts file implements a strategy pattern for authentication. It tries each auth strategy in sequence:
1. JWT Strategy - checks Authorization: Bearer <token> header
2. API Key Strategy - checks X-API-Key header
The first strategy that succeeds sets req.user and calls next(). If all fail, it returns 401 Unauthorized.
Finished in 3ms
AGENT
claude-sonnet-4-5-20250929
Ask a question. @ for files. / for commands

Ứng dụng AI Agent - Computer Use Agent

- **Agent:** OpenClaw, Claude 3.5 Computer Use
- **Environment:** Giao diện hệ điều hành máy tính để bàn, các ứng dụng phần mềm.
- **Observation Space:** Ảnh chụp màn hình, cây giao diện (UI/Accessibility tree), hành động lịch sử,...
- **Action Space:** Điều khiển bàn phím và chuột như click, gõ phím, kéo thả, dùng phím tắt.



OpenAI Universe. (OpenAI, 2016)
OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments, (Xie et al. 2024)

Ứng dụng AI Agent - Robot Agent

- **Agent:** Robot được tích hợp mô hình AI nhúng (Embodied AI)
- **Environment:** Không gian ba chiều của thế giới thực (Nhà máy, văn phòng, đường phố)
- **Observation Space:** Tín hiệu từ cảm biến vật lý, dữ liệu hình ảnh từ hệ thống camera.
- **Action Space:** Điều khiển động cơ học như di chuyển, xoay khớp, cầm nắm vật thể.

Ứng dụng AI Agent - Bài toán VideoQA

Video Toolkit
Tool Name
Description
Input Types
Output Types
Metadata
Tool Card
Visible Frames
Frame 1:...
Frame 80:...
Frame 176:...
Visible Frame Dictionary
In this video, by using a tractor, how many times faster is the cultivation of the same land compared to manual labor?
Calling temporal-grounding-tool to ground the tractor.
Temporal-grounding-tool: Tractor appears from frame <170> to <190>.
Calling image-caption-tool for the 20 frames from <170> to <190>.
Image-caption-tool: ... <176>: The image is composed of two pictures. In the first picture, two people are working together to plow the field, while in the second picture, a single person is operating the tractor. The texts in the image indicate the time it takes to complete the task...
Calling frame-selector to select key frames.
Frame-selector: Look into all current visible frames, frame <176> selected.
Frame <176>:
It seems there are some texts in the image. Calling OCR-tool for frame <176>.
OCR-tool: "45 min" and "45 sec" in frame <176>.
Calling summarizer to summarize all the information.
Summarizer: Using a tractor is 10 times faster than manual labor.

- **Agent:** LLM Planner (ví dụ: GPT-4o) điều phối một bộ công cụ thông qua khung lập luận STAR (Spatiotemporal Reasoning).
- **Environment:** Các video dài chứa sự thay đổi liên tục về diễn biến và hình ảnh.
- **Observation Space:** Câu hỏi đầu vào, từ điển các khung hình đang hiển thị (Visible Frame Dictionary), và kết quả trả về từ công cụ.
- **Action Space:** Luân phiên gọi các công cụ Thời gian (cắt đoạn, chọn keyframe) và Không gian (phát hiện vật thể, OCR, phóng to).

Tool-Augmented Spatiotemporal Reasoning for Streamlining Video Question Answering Task

Ứng dụng AI Agent - Bài toán QA trong Lifelog Retrieval

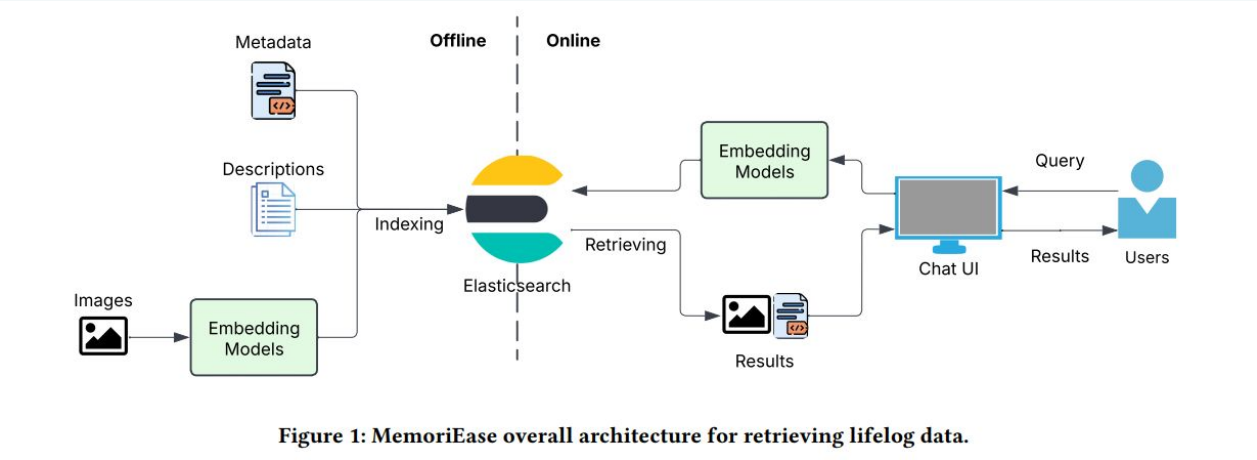


Figure 1: MemoriEase overall architecture for retrieving lifelog data.

MemoriEase 2.0: A Conversational Lifelog Retrieve System for LSC'24 | Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge

MemoriEase 3.0: A RAG-Enhanced Conversational Lifelog Retrieval System at LSC'25 | Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge

- **Agent:** Conversational Search Agent kết hợp với RAG.
- **Environment:** Khối dữ liệu hình ảnh egocentric khổng lồ, lưu trữ dưới dạng vector (CLIP, BLIP-2) và metadata có cấu trúc trên Elasticsearch.
- **Observation Space:** Lịch sử trò chuyện nhiều lượt, tập hợp các ảnh thô truy xuất được, điểm số tương đồng cosine, và captions.
- **Action Space:** Chuẩn hóa và trích xuất bộ lọc trên metadata, tính toán lại vector trọng số từ phản hồi người dùng và kích hoạt quy trình Reranking - Reader để sinh đáp án.

04

Kết luận

Hướng phát triển tương lai

- Lập kế hoạch dài hạn (Advanced Planning)
- Tổng hợp dữ liệu đa nguồn (Data Synthesis)
- Học liên tục (Continuous Learning)
- An toàn hệ thống (AI Safety)
- Giao diện tác tử - máy tính (Agent-Computer Interface)

Hướng phát triển tương lai: lập kế hoạch

- Làm thế nào để thực hiện lập kế hoạch phân cấp? Liệu chỉ cần yêu cầu một LLM, hay còn có điều gì khác cần làm?
- (Tìm kiếm) trong cây có thể đi xa đến đâu?
- Làm thế nào để đánh giá các kế hoạch (một phần)? Hàm giá trị? Mô hình phần thưởng?
- Làm thế nào để thực hiện lập kế hoạch suy đoán dựa trên mô hình cho các agent tổng quát?

Hướng phát triển tương lai: dữ liệu tổng hợp

- Các agent cần học khả năng **cảm nhận-quyết định-thực thi**
- Dữ liệu trên Internet chủ yếu là các sản phẩm từ những quá trình này, không phản ánh chính các quá trình đó
- Tổng hợp dữ liệu với LLMs mang lại khả năng khám phá (một phần) những quá trình ẩn này

Hướng phát triển tương lai: học liên tục

- Hiện tại, lĩnh vực này đang chuyển từ việc sử dụng lệnh điều khiển sang việc sao chép hành vi / tinh chỉnh có giám sát
- Sao chép hành vi có thể không đủ đối với các agent tổng quát; chúng cần phải khám phá các môi trường và học từ **thử và sai**
- Thách thức từ không gian hành động mở, mô hình phần thưởng và an toàn

Hướng phát triển tương lai: an toàn

- Nghiên cứu về an toàn agent vẫn đang tụt hậu so với sự phát triển và triển khai agent
- agent ngôn ngữ
  - Mang tất cả các rủi ro an toàn của LLMs (ví dụ: thiên vị, công bằng, ảo giác, quyền riêng tư, minh bạch)
  - Tăng cường một số rủi ro trong số đó (ví dụ: mất việc làm)
  - Và mang đến nhiều rủi ro mới (ví dụ: hành động không thể đảo ngược)

Hướng phát triển tương lai: giao diện agent-máy tính

- Giao diện máy tính con người -> giao diện agent-máy tính + giao diện agent-con người?
- Hợp tác giữa con người và agent
  - Cách tốt nhất để AI và con người làm việc cùng nhau là gì?
  - Devin vs Cursor vs Github Copilot
- Hầu hết các bài kiểm tra agent giả định cấu hình tự động: loại bỏ yếu tố con người
- Thiết kế ACI được lấy cảm hứng từ thiết kế HCI

Q
A